JASA ARTICLE

A harmonic-cancellation-based model to predict speech intelligibility against a harmonic masker

Luna Prud'homme,[1,a)] Mathieu Lavandier,[1] and Virginia Best[2,b)]
[1]*Univ Lyon, ENTPE, Laboratoire Génie Civil et Bâtiment, Rue Maurice Audin, 69518 Vaulx-en-Velin, France*
[2]*Department of Speech, Language and Hearing Sciences, Boston University, 635 Commonwealth Avenue, Boston, Massachusetts 02215, USA*

ABSTRACT:
This work aims to predict speech intelligibility against harmonic maskers. Unlike noise maskers, harmonic maskers (including speech) have a harmonic structure that may allow for a release from masking based on fundamental frequency (F0). Mechanisms, such as spectral glimpsing and harmonic cancellation, have been proposed to explain F0 segregation, but their relative contributions and ability to predict behavioral data have not been explored. A speech intelligibility model was developed that includes both spectral glimpsing and harmonic cancellation. The model was used to fit the data of two experiments from Deroche, Culling, Chatterjee, and Limb [J. Acoust. Soc. Am. **135**, 2873–2884 (2014)], in which speech reception thresholds were measured for stationary harmonic maskers varying in their F0 and degree of harmonicity. Key model parameters (jitter in the masker F0, shape of the cancellation filter, frequency limit for cancellation, and signal-to-noise ratio ceiling) were optimized by maximizing the correspondence between the predictions and data. The model was able to accurately describe the effects associated with varying the masker F0 and harmonicity. Across both experiments, the correlation between data and predictions was 0.99, and the mean and largest absolute prediction errors were lower than 0.5 and 1 dB, respectively.
 https://doi.org/10.1121/10.0002492

(Received 26 June 2020; revised 8 October 2020; accepted 19 October 2020; published online 30 November 2020)
[Editor: Matthew J. Goupell] Pages: 3246–3254

I. INTRODUCTION

Speech is a complex acoustic signal that has a harmonic structure and a fundamental frequency (F0) that varies around one mean value for a particular talker. Several studies previously showed that when a speech target is masked by a competing sound that is also harmonic, F0 differences (ΔF0) between the target and masker can improve target intelligibility (Brokx and Nooteboom, 1982; Deroche *et al.*, 2014; Leclère *et al.*, 2017). More broadly, it has also long been assumed that differences in voice characteristics, including F0, are critical for selectively attending to one talker in a mixture of talkers (Cherry, 1953).

The mechanisms responsible for the beneficial effects of periodicity in speech-on-speech situations are still not completely understood. This is partly because in such situations, there are two kinds of masking present, and ΔF0 may help to alleviate one or the other or both of them. Energetic masking (EM) refers to a decrease in intelligibility when the target and masker signals overlap in time and frequency such that the target becomes less audible. Informational masking (IM) refers to more central factors that can limit speech intelligibility even when the target is sufficiently audible, such as an inability to segregate the two signals or maintain selective attention to the target speech (see the review in Kidd and Colburn, 2017). IM is typically observed when the masker is very similar to the target or otherwise highly distracting.

To simplify the speech-on-speech problem, a number of studies investigated ΔF0 effects using nonspeech harmonic complex maskers which cause EM but little to no IM. Several mechanisms have been proposed to explain ΔF0 benefits under these conditions. Deroche *et al.* (2014) suggested that listeners could glimpse the target energy in the spectral dips of the masker, which occur between the resolved partials. They showed that speech reception thresholds (SRTs) were better for maskers that provided larger or more numerous spectral glimpses. Another mechanism relies on the idea that listeners are able to detect the harmonic structure of the masker and suppress it when its F0 is different from that of the target. In a study by de Cheveigné *et al.* (1997), listeners were presented pairs of vowels that were either harmonic or inharmonic. They were asked to report the vowels they heard, and each vowel in the pair was scored separately. The overall identification rate was better when the masker was harmonic, but there was no effect of the harmonicity of the target. These results suggested that harmonicity in the masker was more important than harmonicity in the target, which supported the hypothesis of harmonic cancellation over harmonic enhancement. Deroche and Culling (2011) measured SRTs for sentences against harmonic complex tones and investigated the effect of altering the harmonicity of the target or masker using F0

[a)]Electronic mail: luna.prudhomme@entpe.fr, ORCID:0000-0002-8195-6834.
[b)]ORCID:0000-0002-5535-5736.

https://doi.org/10.1121/10.0002492

modulation and reverberation. Their results suggested that masker harmonicity is important to F0 segregation, whereas target harmonicity is of little importance. Steinmetzger and Rosen (2015) measured SRTs for target speech with different degrees of periodicity (using different types of vocoders to create the stimuli) against speech-shaped noise and harmonic complexes with dynamic F0 contours (extracted from speech). The masker envelopes were either stationary or amplitude modulated. Consistent with previous studies, they found that periodicity in the target speech was of little importance to intelligibility but periodicity in the masker could improve SRTs by up to 11 dB.

While several computational models are available that can predict the intelligibility of speech in various kinds of noise, to our knowledge, no model has been shown to predict effects of harmonicity and F0 segregation. Steinmetzger *et al.* (2019) used existing speech intelligibility models to try to predict the data from Steinmetzger and Rosen (2015). They tested four intelligibility models: the extended speech intelligibility index (ESII; Rhebergen *et al.*, 2006), the short-time objective intelligibility (STOI) measure (Taal *et al.*, 2011), the multi-resolution speech-based envelope power spectrum model (mr-sEPSM; Jørgensen *et al.*, 2013), and the correlation-based version of the mr-sEPSM (speech-based envelope power spectrum model, sEPSM^{corr}; Relaño-Iborra *et al.*, 2016). Those four models were chosen because they rely on different theoretical assumptions. The ESII is a signal-to-noise ratio (SNR)-based model that averages the speech intelligibility index (SII) within small timeframes with an implementation of a forward masking function. STOI computes the correlation between clean and noisy speech within auditory filters. mr-sEPSM computes, in temporal windows, the ratio of the envelope powers of the noisy speech and the noise after each are passed through auditory and modulation filtering. The sEPSM^{corr} model is a combination model in which the clean and noisy speech are passed through auditory processing (like in mr-sEPSM) and then the correlation (like in STOI) between the outputs of each auditory and modulation filter is computed. None of the tested models could accurately predict the results from Steinmetzger and Rosen (2015). In particular, all of the models underestimated the benefit due to the masker periodicity. Out of the four models, the best performance was achieved using a modified version of the sEPSM^{corr}, which still underestimated the benefit of masker periodicity by about 5 dB. Steinmetzger *et al.* (2019) suggested that the performance of the models could be improved by implementing a mechanism of enhanced stream segregation dependent on masker periodicity.

The aim of the present study was to develop an intelligibility model able to predict ΔF0 effects in the presence of a harmonic masker by extending an existing SNR-based model to include a harmonic-cancellation component. As a first step toward the prediction of ΔF0 effects, we aimed to develop a model that could accurately predict speech intelligibility against simpler stimuli than those used in Steinmetzger and Rosen (2015). We focused on stationary complex tone maskers with a fixed F0 and no amplitude modulation. This limited the factors we needed to include in this first implementation but still represented an intermediate step between noise and more complex maskers such as speech. It also avoided possible interactions between the effects of masker periodicity and amplitude modulation (Leclère *et al.*, 2017), which could complicate the evaluation of the model.

II. MODEL STRUCTURE

The model presented here is based on a simplified (monaural) version of the (binaural) SNR-based model initially proposed by Collin and Lavandier (2013) and further tested by Vicente and Lavandier (2020). The inputs to the model are the target and masker stimuli at the ears of the listener. The model is composed of two parts: a basic component and a harmonic cancellation component (respectively, black and gray parts in Fig. 1).

The basic component consists of four steps: (1) the signals are passed through a gammatone filterbank (Patterson *et al.*, 1987) with two filters per equivalent rectangular bandwidth, (2) the long-term SNR is computed in each frequency band, (3) weightings are applied according to the SII (ANSI

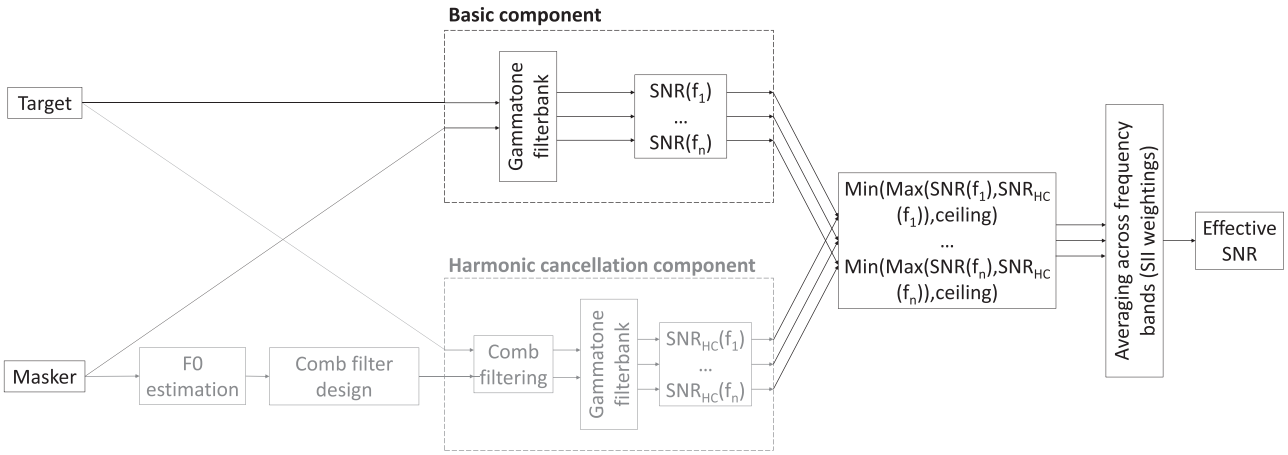

FIG. 1. Structure of the speech intelligibility model with a harmonic cancellation component.

S3.5, 1997), and (4) the SNRs are averaged across frequency bands to obtain an effective SNR.

The harmonic cancellation part of the model is implemented in parallel to the basic component. It consists of three additional steps at the front end: (1) estimation of the F0 of the masker, (2) design of a comb filter that cancels the energy at the estimated masker F0 and its harmonics, and (3) application of this comb filter to both the target and masker signals.

Those two signals are then passed through the gammatone filterbank. The SNR is computed in each frequency band as in the basic component of the model. SNRs with and without harmonic cancellation are computed in parallel, and the higher of the two SNRs is chosen in each frequency band with the assumption that harmonic cancellation is applied only when it is beneficial to intelligibility.

Four parameters are introduced at different steps of the model: a jitter in the F0 estimation, a parameter controlling the shape of the comb filter (the width of its notches), a frequency limit up to which harmonic cancellation is applied, and a ceiling to limit the highest SNR computed by the model. The rationale for the F0 jitter parameter is to simulate imperfections in the estimation of the F0 and the cancellation process. Model versions without jitter, with a fixed jitter, and with a jitter increasing with the F0 of the masker were compared. Different ways to implement the comb filter were tested: one was based on the time-domain comb filter proposed by de Cheveigné (1993), and the other was a frequency-domain filter in which the width and shape of the notches can be modified. A frequency limit up to which harmonic cancellation can be used was investigated, motivated by the idea that spectral components are only resolved by the auditory system within a limited range. The ceiling parameter used when computing the SNR was already present in the model of Collin and Lavandier (2013); several values were tested here to investigate potential interactions with other parameters. Table I summarizes the parameters and their values tested in this study.

III. EXPLORATION OF THE MODEL

A. Behavioral data

Deroche *et al*. (2014) conducted two experiments that measured speech intelligibility against stationary harmonic complex tones with different nominal F0s and different degrees of harmonicity. These two experiments were chosen to test the proposed model that includes harmonic cancellation. Harmonic complex maskers are convenient as they allow an evaluation of the energetic effects of F0 (both spectral glimpsing and harmonic cancellation) in the absence of any significant amount of IM. Sixteen listeners performed the two experiments in the same order. In each experiment, the target stimuli were IEEE (Institute of Electrical and Electronics Engineers) Sentences, and SRTs were measured adaptively using lists of ten sentences.

In experiment 1, SRTs were measured in eight conditions. Maskers were harmonic or inharmonic complex tones with different F0s: 50, 100, 200 and 400 Hz. Inharmonic complex tones were created by randomly jittering (between $\pm$F0/2) each partial from its harmonic position. For each masker type, the SRTs were measured for two conditions: frozen (the same masker was used throughout one block) or fresh (the masker was changed for each sentence). As there was no significant difference between the two conditions, the results presented here were averaged across frozen and fresh conditions. Figure 2 (top, black symbols) shows the mean data reported by Deroche *et al*. (2014). The key results are (1) SRTs decrease with increasing masker F0, (2) harmonic maskers cause less masking than inharmonic maskers, and (3) the difference between harmonic and inharmonic maskers decreases with increasing masker F0. To explain the first result, Deroche *et al*. showed that for both harmonic and inharmonic signals, the widths of spectral dips increase more than the widths of spectral peaks with increasing F0, resulting in more spectral glimpsing opportunities as the masker F0 increases. The second result suggests that there is an advantage linked to the harmonicity of the masker, which could theoretically be achieved via harmonic cancellation (or another harmonicity-based mechanism). The third result can be explained by the fact that for a given F0, spectral dips are wider for inharmonic than for harmonic maskers, and this difference increases with F0. Thus, compared to the harmonic masker, there would be more glimpsing opportunities in the inharmonic masker but less harmonic cancellation.

In experiment 2, SRTs were measured using four types of maskers with various degrees of harmonicity: a harmonic complex, an inharmonic complex created by jittering every partial of a harmonic complex, an inharmonic complex with its first two partials fixed at their harmonic positions, and an

TABLE I. Summary of the parameters and the values tested. The final values are marked in bold.

Parameter	Values
Jitter in masker F0	Fixed: 5–10 Hz Proportional to F0: 0%; 10%; 15%; 20%; **25%**; 30%
Shape of the comb filter	Temporal comb filter Fixed width of notches: 5–10 Hz Width proportional to F0: 0.3; 0.4; 0.5; **0.6**; 0.7F0
Upper frequency limit for harmonic cancellation	Depending on the F0 Fixed: 1000; 2000; 3000; 4000; **5000**; 6000; 7000; 8000; 9000; 10000 Hz; no limit
SNR ceiling	20; 30; **40**; 50 dB

https://doi.org/10.1121/10.0002492

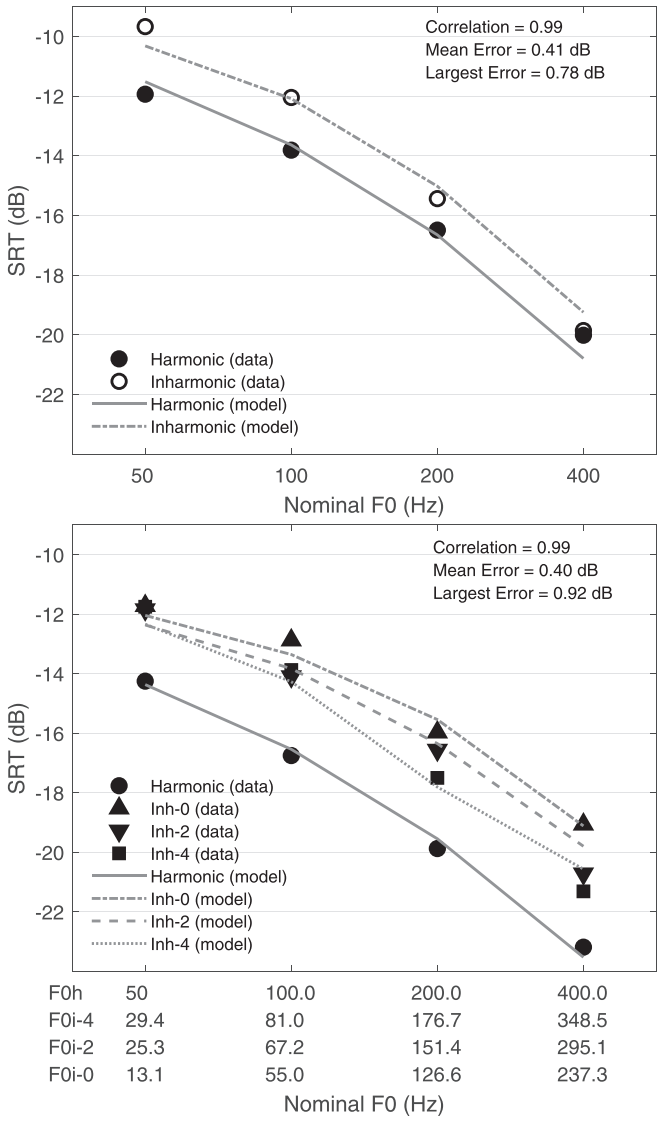

FIG. 2. Mean SRTs measured by Deroche *et al.* (2014; black symbols) and the corresponding model predictions (gray lines) for experiment 1 (top) and experiment 2 (bottom). SRTs are shown as a function of masker F0 for harmonic and inharmonic maskers. For experiment 2, Inh-0 corresponds to the completely inharmonic masker while Inh-2 and Inh-4 correspond to inharmonic maskers with, respectively, the first two and four partials fixed. Those signals had different nominal F0s (as listed) chosen such that the opportunity for spectral glimpsing was similar to that of the corresponding harmonic masker. The parameter values of the model used to generate these predictions were: jitter, 0.25F0; comb filter width, 0.6F0; frequency limit, 5000 Hz; ceiling, 40 dB.

inharmonic complex with its first four partials fixed at their harmonic positions. The harmonic complex had a F0 of 50, 100, 200 or 400 Hz. The inharmonic complexes were based on harmonic complexes at F0s that would create equivalent spectral glimpsing opportunities as the harmonic complex (as measured by a metric proposed by Deroche *et al.*, 2014). Figure 2 (bottom, black symbols) shows the results from experiment 2. As in experiment 1, the SRTs decrease with increasing F0. Moreover, SRTs for the inharmonic complexes with the first two and four partials fixed are lower than for the completely inharmonic masker but higher than those for the harmonic complex. This suggests that these lower partials are useful but all partials are needed to take full advantage of the periodicity of the masker.

B. Application of the model

The model described in Sec. II was optimized using the data from the two experiments described in Sec. III A. Stimuli from those experiments were used as input. Specifically, the target input was composed of 50 target sentences, concatenated to form a single target signal. The masker inputs were 160 realizations of the harmonic maskers used in the experiments; predictions were averaged across realizations. For simplicity, the masker F0 required as model input was not calculated from the signals but taken directly from the original publication of Deroche *et al.* (2014).

Effective SNRs obtained with the model can be compared to SRTs by inverting their signs so that a low SNR corresponds to a high SRT. The model should only be used to predict *differences* in SRTs across conditions in an experiment. To compare predicted and measured SRTs, a reference needs to be chosen (Lavandier *et al.*, 2012). In this study, the reference chosen for each experiment was the average SRT across conditions as done by Collin and Lavandier (2013).

C. Model predictions

Experiment 1 was first used to test the different parameters of the model and narrow down the set of parameter values (see Sec. III D). The resulting parameter combinations were then tested on experiment 2 to choose an optimal combination.

Figure 3(A) shows the predictions of the model for the data in experiment 1 when using only the basic component, which includes spectral glimpsing but not harmonic cancellation. The model predictions do not capture the main effect of harmonicity, confirming that spectral glimpsing alone is not sufficient to explain the results. However, the model predictions do indicate that there are more spectral glimpsing opportunities as the F0 of the masker increases, especially for inharmonic maskers, confirming the explanation provided by Deroche *et al.* (2014). The rest of the panels in Fig. 3 show various unsuccessful implementations of the model that will be used for illustration in Sec. III D.

The final predictions generated by the fully optimized version of the model are shown in Fig. 2 (gray lines) along with the behavioral data from experiment 1 (left) and experiment 2 (right). The model accurately predicts the decrease in SRTs with increasing masker F0 as well as the release from masking associated with harmonicity. It also predicts the difference in SRTs between maskers with different degrees of harmonicity in experiment 2.

The performance of the model was evaluated using the mean absolute prediction error, which corresponds to the mean across conditions of the absolute difference between the behavioral data and the prediction, the largest absolute prediction error, and the Pearson's correlation between the data and the predictions. For both experiments, the Pearson's correlation between data and predictions is 0.99,

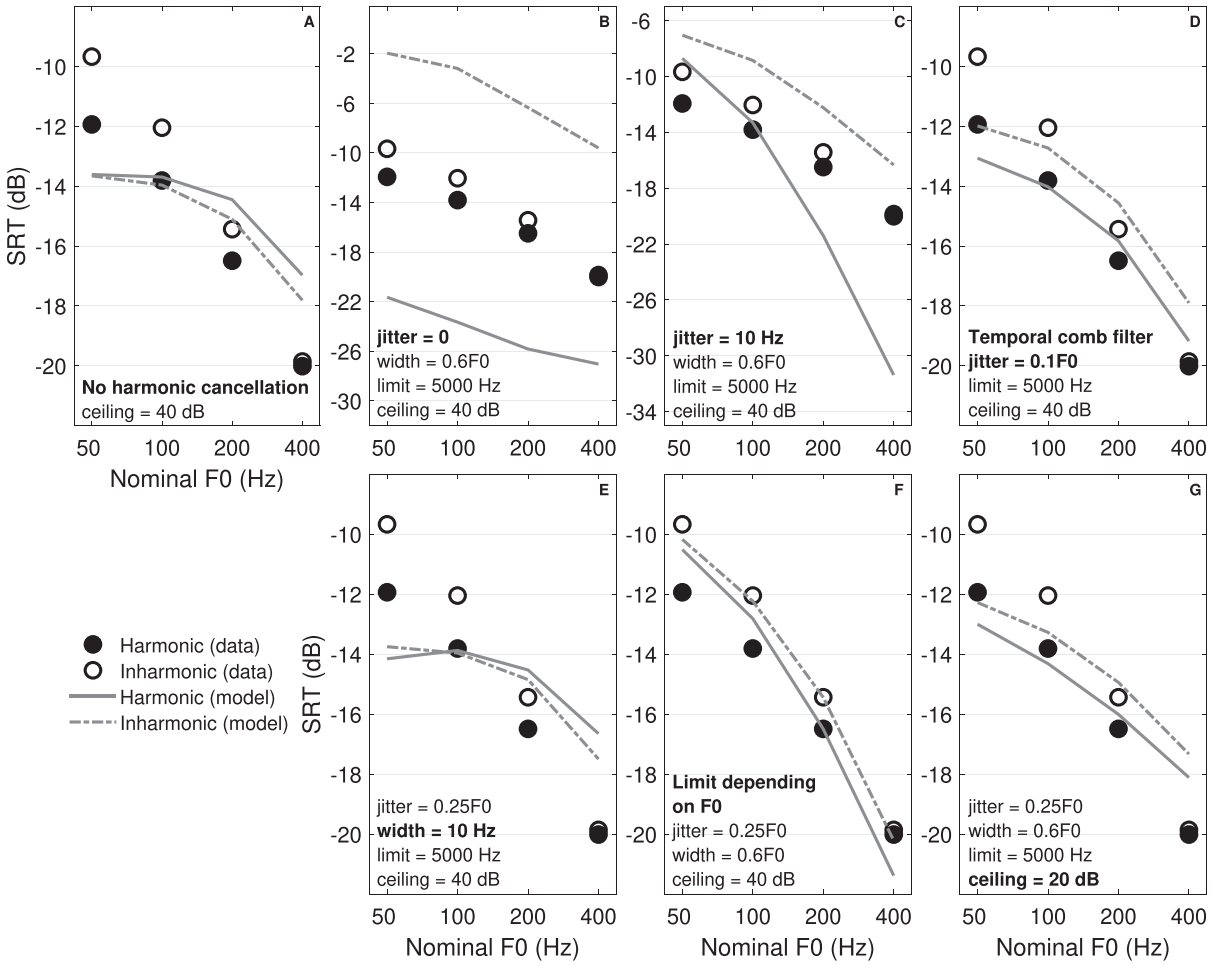

FIG. 3. Mean SRTs measured by Deroche *et al.* (2014, experiment 1) and predictions of various unsuccessful implementations of the model. The parameters were set to the default values unless specified (jitter = 0.25F0, comb filter width = 0.6F0, frequency limit = 5000 Hz, ceiling = 40 dB). (A) Using only the basic component of the model (no harmonic cancellation), (B) in the absence of any jitter in the F0 estimation, (C) using a fixed jitter in the F0 estimation, (D) using a temporal comb filter, (E) using a fixed width for the notches of the comb filter, (F) using a frequency limit depending on F0 (Shackleton and Carlyon, 1994; see text for details), and (G) with a ceiling of 20 dB. Note that the ordinate is different in (B) and (C).

the mean error is lower than 0.5 dB, and the largest error is lower than 1 dB.

As explained in Sec. II, four model parameters needed to be investigated in order to obtain the final predictions. After extensive evaluation of the different parameters and their combinations, for which just some examples are given in Sec. III D and shown in Fig. 3, the following parameter values were selected: jitter 25% (i.e., 0.25F0), width of the comb filter notches 0.6F0, frequency limit 5000 Hz, and ceiling 40 dB. While other combinations of parameters were also successful for predicting the data from experiment 1 of Deroche *et al.* (2014), they were less accurate when tested on experiment 2.

D. Parameter analysis

Different versions of the model and different parameters values were tested to predict the behavioral data of experiment 1 from Deroche *et al.* (2014). A first analysis aimed to set a reasonable range of values for each parameter. Then, different combinations of parameter values were tested to highlight any potential interactions between the parameters.

1. Introduction of a F0 jitter

Figure 3(B) shows the model predictions for experiment 1 when the F0 estimation is assumed to be perfect (i.e., without introducing any jitter in this estimation). The difference between the SRTs for the harmonic and inharmonic conditions at 50 Hz predicted by the model is about 22 dB, which is 20 dB greater than the difference observed in the data (2 dB). This can be explained by the fact that the harmonic maskers are perfectly cancelled by the comb filter when the F0 is perfectly estimated, whereas a physiological implementation might be less perfect.

The model was then evaluated with the introduction of a jitter in the F0 for the creation of the comb filter. The idea is to introduce a jitter in the F0 to account for the fact that the F0 might not be perfectly estimated by the brain but also for any noise in the cancellation mechanism. One solution could have been to introduce a jitter in each notch of the

https://doi.org/10.1121/10.0002492

comb filter, but introducing a jitter in the F0 was simpler and turned out to be sufficient. The magnitude of this jitter was randomly taken from a normal distribution centered at zero. In the rest of this study, the jitter parameter is defined as the standard deviation of this normal distribution.

To obtain model predictions with the jitter parameter that varies randomly from trial to trial, the model is run several times for each condition using a different realization of the stimuli and a different value of the jitter. On each of these "trials," the jitter parameter takes a different random value and produces a different prediction. These predictions are then averaged across trials to estimate the performance of the model. To determine the minimum number of trials needed to produce consistent predictions, we ran the model using 200–2400 trials with a jitter parameter proportional to the F0 (25%). Model performance stabilized after 800 trials, which is the number of trials used for the parameter analysis in Secs. III D 2–III D 5.

2. Influence of the jitter

The performance of the model was explored by testing two options for defining the jitter value. The jitter could either take a fixed absolute value (in Hz) or it could be proportional to the masker F0. A selection of the results is described here.

Figure 3(C) shows that when the jitter was a fixed value (10 Hz), the predicted difference between harmonic and inharmonic maskers increased dramatically with the masker F0, largely because of a very steep improvement in predicted SRTs for harmonic maskers. This is likely because 10 Hz represents a large variation at 50 Hz, but a small variation at 400 Hz, which allows for near-perfect cancellation in the latter case. This is why the difference between harmonic and inharmonic maskers is very small in the 50 Hz condition but very large in the 400 Hz condition, an effect not seen in the behavioral data.

The model was also explored using a jitter proportional to the masker F0. Several values were tested between 10% and 30% of F0. In the examples given here, this parameter was tested in combination with the parameter representing the width of the notches of the comb filter (see Sec. III D 3). Figure 4 presents the mean and largest prediction errors for five jitter values between 10% and 30% of F0, and three notch width values (0.4F0, 0.5F0, and 0.6F0). Note that the performance of the model is influenced by both the jitter and the width of the notches. Overall, however, better performance is obtained when the jitter value is at least 20%. In this example, the optimal value in which the errors are minimized or reach a plateau is approximately 20% or 25% for the width of the notches at 0.5F0 and 0.6F0, respectively. The final model predictions using a jitter of 25% (i.e., 0.25F0) and a notch width of 0.6 F0 are shown in Fig. 2 (top).

3. Design of the comb filter

Two versions were tested for the comb filter that cancels the energy at F0 and its harmonics. The first option was a simple time domain comb filter proposed by de Cheveigné (1993). The impulse response of this filter is given by $h(t) = \frac{1}{2}(\delta(t) - \delta(t - T))$, where T is defined by $T = 1/\text{F0}$.

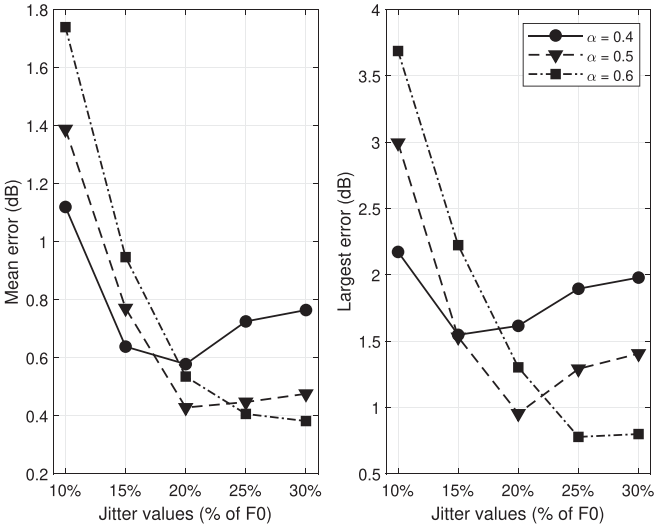

FIG. 4. Mean and largest errors in the predictions for Deroche *et al.* (2014, experiment 1) as a function of the F0-dependent jitter value for three values α of the F0-dependent width of the notches of the comb filter (frequency limit = 5000 Hz, ceiling = 40 dB).

Figure 5 shows the impulse response of such a comb filter (top panel). The predictions of the model using this filter to model harmonic cancellation are displayed in Fig. 3(D). The predictions are better than without any harmonic

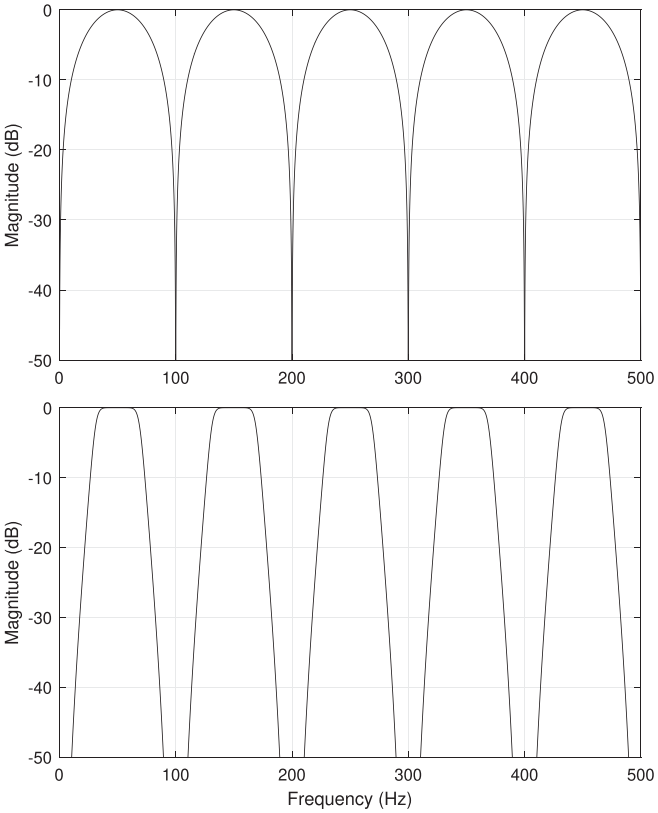

FIG. 5. Example of the frequency response for F0 = 100 Hz of the time domain comb filter (de Cheveigné, 1993) and the comb filter used in the model with a width of notches equal to 0.6F0 (respectively, top and bottom panels).

cancellation [basic component of the model; see Fig. 3(A)], but this model version cannot predict with high accuracy the results of Deroche *et al.* (2014). In particular, the difference between harmonic and inharmonic maskers is underpredicted for lower-F0 maskers. Closer inspection of the model outputs indicated that the filter proposed by de Cheveigné (1993), when implemented with a F0 jitter, is too narrow to have a sufficient effect on the prediction.

Another approach was considered that used an infinite impulse response (IIR) comb filter created in MATLAB (The MathWorks, Natick, MA) using the function fdesign.comb (Fig. 5, bottom panel). This approach allowed us to control the width of the notches and the shape of the filter in order to cancel more of the masker energy. We found that the cancellation was most effective when the width of the notches was proportional to the masker F0. Figure 3(E) shows the predictions of the model when the width of the notches was fixed at 10 Hz. The model performance was significantly worse when the width of the filter was fixed compared to when it was proportional to the masker F0 (Fig. 2). The width of the notches was then defined by width $= \alpha$F0. As shown previously in Fig. 4, the best results were obtained for widths of 0.5F0 or 0.6F0 (depending on the value of the jitter). The final comb filter is similar to the one proposed by de Cheveigné (1993), but it cancels more of the masker energy so that harmonic cancellation is more efficient and produces better predictions of the data.

4. Frequency limit

Another parameter explored was the frequency limit up to which harmonic cancellation is applied. In the frequency bands below that limit, the model is applied as explained in Sec. II (choosing the maximum between the SNR from the basic component and the SNR from the harmonic cancellation component) and in the frequency bands above that limit, only the basic component is applied. The idea here is that harmonic cancellation might not be useful above a certain frequency limit (for example, once the harmonics are unresolved by the auditory system). The frequency limit could either be fixed or depend on the masker F0, but the model performance was generally better when the frequency limit was fixed. Figure 3(F) displays the model predictions when harmonic cancellation was applied only for frequency bands in the region where the harmonics of the masker would be resolved. The limit between resolved and unresolved harmonics was calculated using the definition given by Shackleton and Carlyon (1994). In this case, the frequency limit depends on the F0 of the masker. The partials are considered as resolved when fewer than two partials pass through the 10-dB bandwidth of an auditory filter and unresolved when there are more than 3.25 partials per filter. The performance of the model is not satisfactory in this case as the difference between the SRTs of harmonic and inharmonic maskers is reduced at low F0s and increased at high F0s.

Fixed frequency limits between 1000 and 10 000 Hz were tested. The model performance was poorest when the limit was below 3000 Hz. For limits of 3000 Hz and greater, the model performance was consistently good with optimum performance at 5000 Hz (Fig. 6).

5. SNR ceiling

The ceiling parameter was introduced and tested in previous versions of the model that do not include harmonic cancellation (Collin and Lavandier, 2013; Vicente and Lavandier, 2020). Ceiling represents the highest value that the SNR can take in each frequency band. In Vicente and Lavandier (2020), it was fixed at 20 dB and was necessary for accurate predictions in the presence of amplitude modulated noise maskers, where the SNR can approach infinity in the dips of the masker. While it was not clear that this parameter would be essential in the present model given that the signals are stationary, it could be important for limiting the SNR in cases where applying the comb filter greatly reduces the masker energy. After an exploration of different ceiling values (from 20 to 50 dB), an optimal value of 40 dB was chosen. The reason for this high value appears to be that the maskers used by Deroche *et al.* (2014) were not speech shaped but rather had a flat spectrum, which results in very high SNRs in certain frequency bands. Figure 3(G) shows the predictions of the model when the ceiling was at 20 dB. This lower ceiling appears to prevent the accurate prediction of spectral glimpsing effects.

IV. DISCUSSION

The present paper describes the implementation of harmonic cancellation in a SNR-based speech intelligibility model, as well as the optimization of the new model. The model describes SRTs measured against stationary harmonic and inharmonic complexes with high accuracy. The mean and largest prediction errors were less than 1 dB and similar to those reported previously for other speech intelligibility models predicting SRTs for speech in noise (Beutelmann and Brand, 2006; Collin and Lavandier, 2013; Lavandier *et al.*, 2012). As a comparison, Steinmetzger *et al.* (2019) used four models to predict the masker periodicity benefit and obtained, for the best model, a mean error of about 5 dB. Note that the present study involved simpler stimuli (with a monotonous F0) than those used by Steinmetzger *et al.* (2019). Note also that the proposed model still needs to be tested on data not used to define its parameters, so that its prediction power can be evaluated.

The model does not fully capture some of the effects observed in the behavioral data. In the data (experiment 1), the difference between SRTs for harmonic and inharmonic maskers is reduced with increasing masker F0 to the point that there is no difference for a masker with a F0 of 400 Hz. According to Deroche *et al.* (2014), this effect is due to increased spectral glimpsing opportunities in the inharmonic compared to those in the harmonic masker. In the model

https://doi.org/10.1121/10.0002492

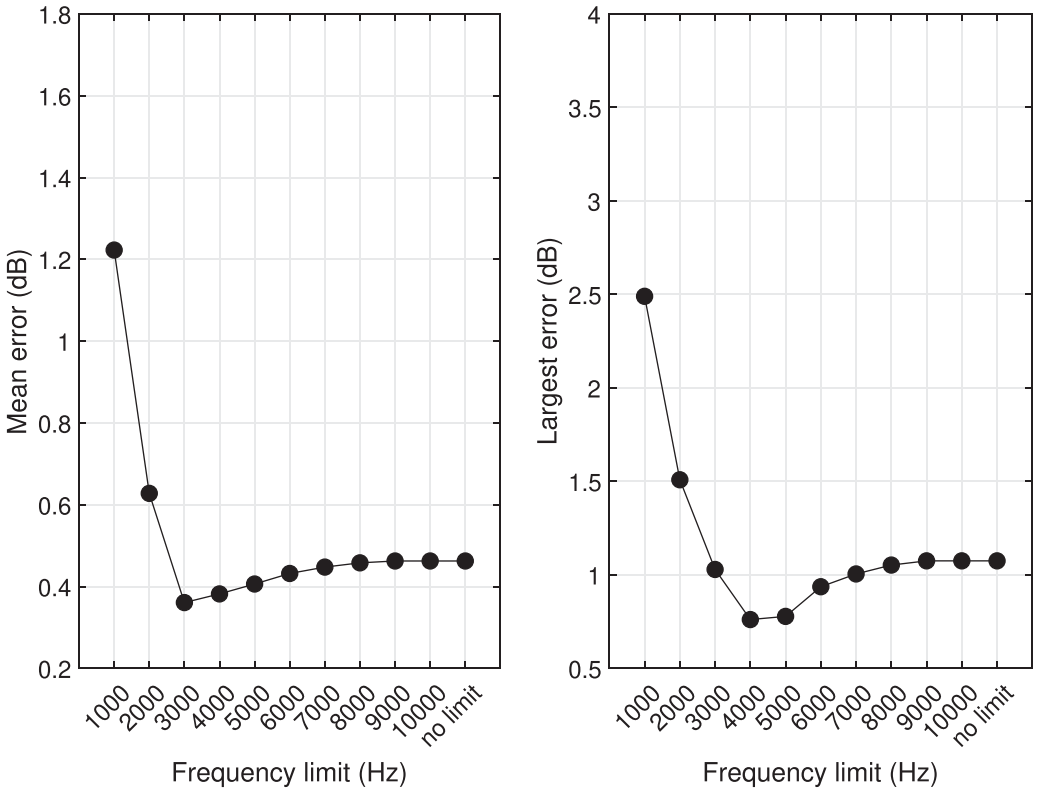

FIG. 6. Mean and largest errors of the predictions for Deroche *et al.* (2014, experiment 1) as a function of the frequency limit up to which harmonic cancellation is applied. The other parameters of the model were: jitter, 0.25F0; comb filter width, 0.6F0; ceiling, 40 dB.

predictions, however, SRTs are still slightly lower for the harmonic masker.

The parameter analysis that investigated whether a frequency limit for harmonic cancellation was important for the performance of the model did not result in a conclusive answer. Putting a frequency limit at 5000 Hz only resulted in a very small advantage compared to other values or running the model without a frequency limit. However, it is worth noting that the harmonic cancellation component of the model is implemented in a way that does not take into account auditory filtering. The comb filter accounting for harmonic cancellation is directly applied to the waveform of the target and masker. Introducing a frequency limit, even though its effect was limited in this experiment, could be important in future applications of the model where the use of harmonic cancellation at higher frequencies produces erroneous predictions. Further work may be needed to determine exactly what the appropriate frequency limit is. In a third experiment, Deroche *et al.* (2014) showed that there was a consistent benefit due to harmonicity in the masker whether the target speech was low-pass, band-pass, or high-pass filtered, which indicates that the mechanism underlying the harmonicity advantage is still active after 2535 Hz (the cut-off frequency for the high-pass filter). The results of our parameter analysis, suggesting a frequency limit close to 5000 Hz, are broadly consistent with their findings.

In the version of the model described here, harmonic cancellation was applied in each frequency band only if it improved the SNR in that band. We also tested a version of the model in which the choice between harmonic cancellation and the basic component was made after averaging the SNR over all frequency bands (i.e., harmonic cancellation was applied in all or none of the frequency bands). The predictions obtained for the experimental data of Deroche *et al.* (2014) were always slightly better when the choice of applying harmonic cancellation or not was made independently in each frequency band. While this is an issue that deserves further investigation, a per-channel decision seems plausible. In their discussion on the implementation of a model of harmonic cancellation, Guest and Oxenham (2019) wrote that "one simple possibility might be to selectively apply the cancellation filter to the outputs of auditory filters that are dominated by the masker (i.e., little representation of the target periodicity is present, or the SNR is poor). Thus, the outputs of auditory filters with a good SNR would be left unaffected while the SNR at outputs of auditory filters with unfavorable SNRs before processing might be improved by cancellation."

The next step in modeling speech intelligibility against harmonic maskers would be to take into account F0 variations over time. The maskers used here had a steady F0, which is an ideal case for harmonic cancellation but is not representative of speech maskers, which have intonation in their F0 pattern (as well as unvoiced parts with no F0). Leclère *et al.* (2017) measured SRTs against both monotonized and intonated harmonic complexes and showed that ΔF0 effects are much reduced when the masker F0 is intonated. This result implies that harmonic cancellation might

be less effective when F0 varies over time and might play a smaller role (or none at all) in those situations. In order to take into account F0 variations, the model would need to operate over shorter time frames, where the F0 to be cancelled may change from frame to frame. An important step in such a modification would be to establish the appropriate duration of these time frames.

Another step forward would be to develop a model that is able to predict benefits of spatial separation and amplitude modulation in addition to harmonic cancellation. This could be approached by adding the better-ear and binaural unmasking components as implemented in Collin and Lavandier (2013). This might not be as straightforward as it seems, given that effects of better-ear listening and amplitude modulation are implemented by computing the SNR in rather short time frames (on the order of 25 ms). In the case of harmonic maskers, such time frames might be too short relative to the period of the masker in order to "see" the spectral peaks and dips in the spectrum. Thus, different time windows for different components of the model might need to be considered.

V. CONCLUSION

A SNR-based speech intelligibility model with an implementation of harmonic cancellation was proposed to take into account changes in EM associated with F0 differences and masker harmonicity. An analysis of the four parameters introduced in the model was made, and values were chosen that optimized the performance of the model. The model was able to accurately predict the data from two experiments in which speech intelligibility was measured against maskers with different F0s and degrees of harmonicity. This work represents a critical step toward a comprehensive model that can predict speech intelligibility in more realistic situations such as those involving competing talkers.

ACKNOWLEDGMENTS

The authors would like to thank Mickael Deroche for sharing his data. This work was performed within the LabEx CeLyA (ANR-10-LABX-0060/ANR-16-IDEX-0005) and funded by the "Fondation Pour l'Audition" (Speech2Ears grant). V.B. was supported, in part, by National Institutes of Health-National Institute on Deafness and Other Communication Disorders (NIH-NIDCD) Award No. DC015760.

ANSI S3.5 (**1997**). *Methods for Calculation of the Speech Intelligibility Index* (Acoustical Society of America, New York).

Beutelmann, R., and Brand, T. (**2006**). "Prediction of speech intelligibility in spatial noise and reverberation for normal-hearing and hearing-impaired listeners," J. Acoust. Soc. Am. **120**, 331–342.

Brokx, J. P., and Nooteboom, S. G. (**1982**). "Intonation and the perceptual separation of simultaneous voices," J. Phonetics **10**, 23–36.

Cherry, E. C. (**1953**). "Some experiments on the recognition of speech, with one and with two ears," J. Acoust. Soc. Am. **25**, 975–979.

Collin, B., and Lavandier, M. (**2013**). "Binaural speech intelligibility in rooms with variations in spatial location of sources and modulation depth of noise interferers," J. Acoust. Soc. Am. **134**, 1146–1159.

de Cheveigné, A. (**1993**). "Separation of concurrent harmonic sounds: Fundamental frequency estimation and a time-domain cancellation model of auditory processing," J. Acoust. Soc. Am. **93**, 3271–3290.

de Cheveigné, A., McAdams, S., and Marin, C. M. H. (**1997**). "Concurrent vowel identification. II. Effects of phase, harmonicity, and task," J. Acoust. Soc. Am. **101**, 2848–2856.

Deroche, M. L. D., and Culling, J. F. (**2011**). "Voice segregation by difference in fundamental frequency: Evidence for harmonic cancellation," J. Acoust. Soc. Am. **130**, 2855–2865.

Deroche, M. L. D., Culling, J. F., Chatterjee, M., and Limb, C. J. (**2014**). "Speech recognition against harmonic and inharmonic complexes: Spectral dips and periodicity," J. Acoust. Soc. Am. **135**, 2873–2884.

Guest, D. R., and Oxenham, A. J. (**2019**). "The role of pitch and harmonic cancellation when listening to speech in harmonic background sounds," J. Acoust. Soc. Am. **145**, 3011–3023.

Jørgensen, S., Ewert, S. D., and Dau, T. (**2013**). "A multi-resolution envelope-power based model for speech intelligibility," J. Acoust. Soc. Am. **134**, 436–446.

Kidd, G., Jr., and Colburn, H. S. (**2017**). "Informational masking in speech recognition," in *The Auditory System at the Cocktail Party*, edited by J. C. Middlebrooks, J. Z. Simon, A. N. Popper, and R. R. Fay (Springer Nature, New York), pp. 75–109.

Lavandier, M., Jelfs, S., Culling, J. F., Watkins, A. J., Raimond, A. P., and Makin, S. J. (**2012**). "Binaural prediction of speech intelligibility in reverberant rooms with multiple noise sources," J. Acoust. Soc. Am. **131**, 218–231.

Leclère, T., Lavandier, M., and Deroche, M. L. (**2017**). "The intelligibility of speech in a harmonic masker varying in fundamental frequency contour, broadband temporal envelope, and spatial location," Hear. Res. **350**, 1–10.

Patterson, R., Nimmo-Smith, I., Holdsworth, J., and Rice, P. (**1987**). "An efficient auditory filterbank based on the gammatone function," in *Present. Inst. Acoust. Speech Group Audit. Model. R. Signal Res. Establ.*, Vol. 34.

Relaño-Iborra, H., May, T., Zaar, J., Scheidiger, C., and Dau, T. (**2016**). "Predicting speech intelligibility based on a correlation metric in the envelope power spectrum domain," J. Acoust. Soc. Am. **140**, 2670–2679.

Rhebergen, K. S., Versfeld, N. J., and Dreschler, W. A. (**2006**). "Extended speech intelligibility index for the prediction of the speech reception threshold in fluctuating noise," J. Acoust. Soc. Am. **120**, 3988–3997.

Shackleton, T. M., and Carlyon, R. P. (**1994**). "The role of resolved and unresolved harmonics in pitch perception and frequency modulation discrimination," J. Acoust. Soc. Am. **95**, 3529–3540.

Steinmetzger, K., and Rosen, S. (**2015**). "The role of periodicity in perceiving speech in quiet and in background noise," J. Acoust. Soc. Am. **138**, 3586–3599.

Steinmetzger, K., Zaar, J., Relaño-Iborra, H., Rosen, S., and Dau, T. (**2019**). "Predicting the effects of periodicity on the intelligibility of masked speech: An evaluation of different modelling approaches and their limitations," J. Acoust. Soc. Am. **146**, 2562–2576.

Taal, C. H., Hendriks, R. C., Heusdens, R., and Jensen, J. (**2011**). "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," IEEE Trans. Audio Speech Lang. Process. **19**, 2125–2136.

Vicente, T., and Lavandier, M. (**2020**). "Further validation of a binaural model predicting speech intelligibility against envelope-modulated noises," Hear. Res. **390**, 107937.